图书情报工作
第64卷 第14期 2020年7月

知识图谱在实体检索中的应用研究综述*

■ 阮光册[1] 樊宇航[1] 夏磊[2]

[1] 华东师范大学经济与管理学部信息管理系 上海 200241 [2] 上海图书馆 上海 200031

摘 要：[目的/意义]梳理基于知识图谱的实体检索的研究脉络和重点，探索未来该领域的发展方向。[方法/过程]概述基于知识图谱的实体检索的形式化定义、实现路径以及主要的数据源；根据检索任务，将实体检索划分为匹配检索、扩展检索和推荐检索3种实现场景，并对其实现方法进行综述。[结果/结论]随着应用的不断深入，基于知识图谱的实体检索研究开始关注如何优化用户的检索体验和提供多样性的检索结果，未来将在检索结果可解释性、跨领域知识图谱检索等多个方面展开深入的研究。

关键词：实体检索 知识图谱 信息检索

分类号：G250

DOI：10.13266/j.issn.0252-3116.2020.14.013

1 引言

检索是人们获得信息和知识的重要途径。随着技术的发展，检索技术已经从最初由用户提供检索需求、系统返回给用户信息列表，逐步改进到能够主动为用户提供信息推荐的服务方式[1]。然而，用户在使用检索系统获取信息时，有时候希望能够获得汇总的答案以及简要的总结，这个过程即为实体检索。实体检索的特点是直接给出答案或给出有关类别的列表。实体检索对用户来说，信息消费代价相对不高，同时增强了用户的体验感。研究表明[2]，在用户的信息检索行为中，以实体作为检索意图占检索活动的比重达到52%，这使得实体检索成为检索系统必不可少的功能之一。

目前，随着大规模知识图谱（又称知识库）的出现，如YAGO[3-5]、DBpedia[6]、Freebase[7]、NELL[8]、Probase[9]等，方便了检索系统从数据资源中抽取、组织和管理知识。基于知识图谱的实体检索系统主要是利用数据集的结构化特征，通过实体及实体之间的语义关系来表达知识，不仅能够帮助用户检索和发现感兴趣的相关实体，也能够以探索的形式为用户提供深入检索的服务。这种方式从语义层面理解用户意图，借助知识图谱的网络信息实现复杂的关联查询，提升了检索过程中知识的理解层次。

近年来，有关知识图谱的研究取得了极大的进展，关于知识图谱的综述性研究也陆续发表，这些综述性研究包括对知识抽取、知识表示、知识推理的研究[10]，也包括知识图谱构建和知识融合问题的综述[11-12]。然而，在这些研究中，仍缺少对知识图谱在实体检索中的研究进展进行系统梳理的文献。为此，本文将Web of Science核心合集作为主要文献来源，以主题=（“knowledge graph” or “KG”）and “retriev*”为检索式进行检索（检索时间为2019年4月11日）。对于检索出来的结果，首先进行了歧义筛选，将医学领域的文献中表示“恢复”（retrieval）的检索结果剔除，然后将研究领域限定为计算机和信息科学相关，再通过人工阅读去除了主要内容与基于知识图谱的实体检索无太多联系的文献，最后保留81篇文献作为本文的主体素材。

1.1 知识图谱概述及其在检索中的应用

知识图谱旨在描述现实世界中存在的实体以及实体之间的关系，从本质上来看，是一种语义网络。早在20世纪60年代末，M. R. Quillian就提出通过构建复杂的元素网络关联，来模拟人类语言行为的知识基础，利用带有标记的有向图，借助事物属性以及事物之间的语义关系发现与节点有关的知识[13]。随后，R. F.

* 本文系上海市软科学研究计划项目“数据智能时代基于深度学习的情报处理分析研究”（项目编号：19692106600）研究成果之一。

作者简介：阮光册（ORCID：0000-0001-8685-5234），副教授，博士，硕士生导师，E-mail：rgc1976@126.com；樊宇航，硕士研究生；夏磊（ORCID：0000-0002-9141-2689），上海图书馆会展中心副主任，副研究馆员，硕士。

收稿日期：2019-11-06 修回日期：2020-03-05 本文起止页码：126-135 本文责任编辑：易飞

Simmons[14]在进行自然语言问答系统的研究时，将 M. R. Quillian 设计的知识表示模型定义为“语义网络”(semantic network)，认为其是语义分析方法上的初步探索。

语义网络和知识图谱均采用图的形式进行知识表示。但语义网络中节点表示对象和概念，边表示节点之间的关系，节点和边的值可由用户进行定义，这就为多源数据融合带来了困难；而知识图谱中，节点表示实体或者属性值，边表示关系或者属性，将事物的属性以及事物之间的语义关系显式地表示出来，利用三元组的形式进行刻画。为此，知识图谱的构建更加规范，结构简洁直观、使用灵活丰富，也确保了数据的质量。近年来，随着开放链接(linked open data，简称 LOD)等项目[15]的开展，大量 RDF(resource description framework)数据被发布，互联网从文档万维网变成了包含大量描述各种实体和实体之间丰富关系的数据万维网(web of data)。2012 年 5 月，Google 发布了知识图谱，其初衷是为了提高搜索引擎的能力，改善用户的搜索质量以及搜索体验。最初的知识图谱是建立在公开的 RDF 数据源之上的，随着机器学习和数据挖掘方法的应用，人们开始从非结构化的网页数据中自动发现新的实体和实体间的关系[16]。目前，大规模知识图谱库的研究和应用在学术界和工业界引起了足够的注意力，现已被广泛应用于智能搜索、智能问答、个性化推荐、内容分发等领域。

领域本体是形成知识图谱的一个重要步骤，也是知识图谱的基础和骨架。领域本体是描述一个领域的术语集合，其所定义的类、关系、函数、公理和实例，在模式层约束和管理知识图谱。在现有的研究成果中，依托于已有的成熟本体库，基于知识图谱的实体检索在生物医学领域的研究和实践较为深入。将单一[17]乃至多个来源[18-20]的异构数据整合成知识图谱用于检索，是学者们早期关注的重点。随着应用研究的深入，整合知识图谱仅仅用于检索已无法满足用户需求，在检索中发现新的知识，进行知识推理[21-23]，乃至借助数据挖掘[24-25]的方法进行知识发现的研究也开始展开。目前，基于知识图谱的实体检索开始应用在更多的领域，例如空间地理研究[26-27]、农业生产[28-29]等，在其他领域，例如科研[30-31]和企业[32-33]同样关注该技术对于知识管理的应用价值。从整体上看，基于知识图谱的实体检索的应用场景，从最初单纯地构建知识图谱用于检索，发展到之后集成挖掘推理的平台级应用，领域也从成熟的生物医药领域不断扩展，包括与新技术[34]的融合。

1.2 基于知识图谱的实体检索

实体是指现实或虚拟世界中具有特定语义的任何对象或者概念，是知识图谱中的最基本元素。同时，实体间都存在一定的关系，其基本形式主要包括(实体 1 - 关系 - 实体 2)和(实体 - 属性 - 属性值)等。

如果将实体表示为 e，基于知识图谱的实体检索可采用{D，q，R(q，ei)}进行表示。其中 D 是知识图谱库，即知识库，知识库由大量实体及实体之间的关系组成；q 表示为用户的查询请求，R(q，ei)则表示用来度量用户的查询与数据集中实体 ei 的相关性或相似性，并根据评分结果，给出针对 q 的 ei 结果排序。

一般来说，基于知识图谱的实体检索实现路径如图 1 所示，主要分为 2 个部分：知识图谱构建过程和检索实现过程。

知识图谱的构建是一个迭代的过程，每一轮迭代均包含知识抽取、知识融合和知识加工 3 个阶段：①知识抽取是指从多源异构数据源中抽取实体、属性以及实体间的相互关系，在此基础上形成本体化的知识表达；②通过抽取获取的新知识需要进行整合，这个过程主要实现 2 个目标：实体消歧和共指消解，前者为解决同名实体产生歧义的问题，后者则是将指向同一实体的项关联(合并)到唯一正确的实体对象；③知识加工是知识结构化的过程，构建本体是知识加工阶段常采用的方式，即设定类、实例、属性、关系、规则等元素的标准。知识推理和质量评估是知识加工的两个重要过程。其中，知识推理主要目的是建立实体间的新关联，这有助于拓展和丰富知识图谱；质量评估通常与实体对齐的任务同时进行，通过舍弃置信度较低的知识，保留置信度较高的知识，保障知识图谱的数据质量。

在用户检索过程中，核心环节是如何将用户的检索输入与知识图谱进行交互。首先，需要系统将用户输入的自然语言检索式、实体对或 SparQL 语句，转化成查询子图并与整个知识图谱进行匹配。随后，检索系统将识别检索式中的语义实体，并对实体间的结构关系进行查询扩展与推理。最后，系统对检索处理的结果进行相关性的排序，将最能符合用户需求的结果提供给用户。在实施检索过程中，由于检索处理会受到不同匹配策略的影响，同时为了提高检索的效率，系统还需要构建相应的存储方式和索引。具体过程见图 1。

1.3 主要数据集

大规模知识图谱是实体检索的基础。在国外知识

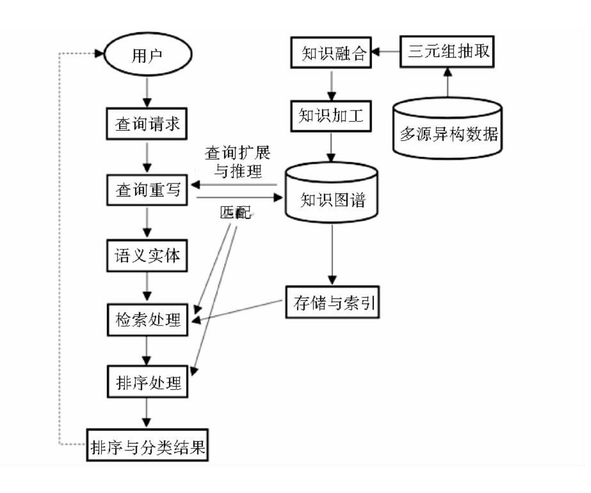


图 1 知识图谱的检索路径

图谱的成熟产品包括:莱比锡大学的 DBpedia、MetaWeb(2010 年被 Google 收购)的 Freebase、Google 的 Knowledge Graph 与 Knowledge Vault、维基媒体基金会的 Wikidata,马克思・普朗克研究所的 YAGO、微软公司的 Probase 和 Bing Satori、沃尔夫勒姆公司的 Wolfram Alpha、Facebook 公司的 Fackbook Social Graph 以及斯坦福大学的 ImageNet。这些知识图谱目前应用在多个领域,如 Google Search Engine、Wikipedia、Apple Siri 或计算机视觉等产品。

国内商业机构在大规模知识图谱的研究上也取得了一定成果,例如百度公司应用于百度搜索的百度知识图谱、应用于搜狗搜索引擎的搜狗知立方等。国内一些研究机构的知识图谱包括复旦大学 GDM 实验室的 CN-Dbpedia、北京大学的 PKU-PIE、清华大学的 XLORE、中国科学院自动化研究所的 Belief-Engine 等。

2 实体检索的应用

基于知识图谱的实体检索,对数据资源组织有更细致的要求,不仅要考虑数据资源的外部特征,更需要通过三元组抽取、知识融合和知识加工这一系列的过程,将数据资源组织成结构化的知识表达形式。在检索过程中,系统更注重"知识"概念本身,而非知识概念外在的形式特征。用户输入检索请求后,系统需要对用户的需求进行语义理解,形成查询子图,实现与知识图谱的匹配计算。最后,检索返回的结果也是相关实体、实体组或者实体关系的三元组。相较于单纯的包含有用户检索式文本片段的检索结果而言,基于知识图谱的实体检索结果提供了更丰富的语义特征和结构化表示形式。

基于知识图谱的检索与基于本体的检索虽有很多的相似之处,但本体是从 schema 的角度来进行定义,而知识图谱更强调数据层的构建,使得基于知识图谱的检索相较于基于本体的检索多了结构相似性、路径距离等图算法中应用到的方法。

张香玲等[35]将实体检索分为相关实体检索和相似实体检索。本文认为,在知识图谱的实体检索应用过程中,研究的重点是如何计算用户非结构化的检索需求与知识图谱网络形式之间的语义匹配。随着实践应用的深入,研究的重点也开始关注如何优化用户的检索体验和提供多样性的检索结果上。为此,在上述研究[35]的基础之上,本文将基于知识图谱的实体检索划分为 3 种任务场景:匹配检索、扩展检索和推荐检索。

2.1 匹配检索

基于知识图谱的实体检索,在匹配算法上,相较于不基于知识图谱的实体检索,除了语义相似性之外,还多了结构相似性,在丰富了语义特征的同时,对匹配计算以及如何对候选实体进行排序提出了新的要求。一般来说,匹配检索的基本思路是:首先为每个实体创建生成文档,将每个实体相关的三元组看作实体的生成文档,计算文档与用户查询之间的相关度,排序后得到候选实体[36-37]。结构性实体模型和层次结构实体模型是 2 种主要的匹配计算模型,模型的对照如图 2 所示:

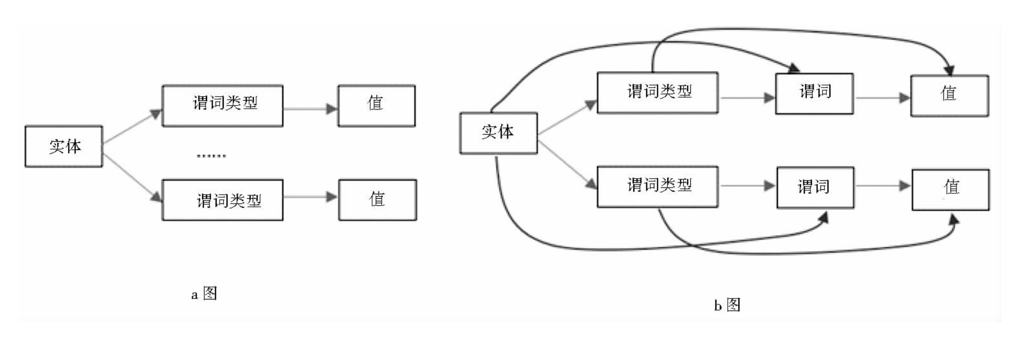


图 2 实体模型对照

图2中,a图为结构性实体模型,b图为层次结构性实体模型。从图中可以看出,结构性实体匹配模型采用谓词结构表示实体,但没有利用到谓词之间的语义信息;而层次结构性实体模型则采用两层结构,通过实体谓词的类型,为检索匹配计算提供了更多的语义特征。实验表明,层次结构性实体模型比无结构化实体模型以及结构化性实体模型的检索效果都要好。

匹配计算的效率问题,是研究者考虑的另一个重要的问题。在众多的研究中,构建索引和数据分区是主要的研究策略。F. Lashkari等[38]混合了反向索引、Treap和wavelet树,为实体、类型和关键字构建了语义混合索引数据结构,提高了检索效率。A. Katib等[39]设计了RIQ工具,通过使用过滤索引,以及每组相似的RDF图的单独索引,实现高效查询。在数据分区方面,Y. Hao等[40]提出一种面向关联的图分区方法Assc,在查询过程中明显减少了各分区之间的交互,从而提升效率。W. Zheng等[41]在天际线查询中,利用处于天际线的实体,来进行数据分区,从而避免一些不必要的分区进入候选集。此外,也有学者直接从候选集修剪方面入手进行优化,通过知识图谱构建摘要,极大地改善了子图匹配的效率[42]。也有学者[43]利用预先存储的路径视图,重写知识图谱中的路径查询,避免了指数级的嵌套循环和路径候选集。

随着大规模知识图谱的广泛应用,学者们发现知识图谱本身存在大量实体关系和属性缺失的问题(如DBpedia),同时,节点之间的关系也存在着不确定性。如何提高匹配检索的准确性,成为研究的重点。解决这一类问题常用的方法是:图的切割以及实体预测。有研究[44]通过迭代,不断地切割不确定图,将不合格的匹配结果剪掉可以在一定程度上提高检索效率。然而,这种方法无法解决不确定图存在的可达性问题。为此,有研究[45]借助概率计算,选择一些不确定的边缘使其概率接近0或者1,以此降低检索时的不确定性。为解决知识图谱的不完整性,研究主要是围绕对缺失部分进行预测,概率语言模型是常采用的方法,通过对缺失的三元组和实体进行预测,来填补知识图谱中缺失的部分[46－47]。然而,在概率计算时,大多假设候选实体和查询关键词分布是条件独立的,忽略了两者在文档中的语义关联。为此,一些研究尝试采用计算实体类别和检索关键词之间的包含度[48],确保候选实体能够包含于结果集中,也有研究引入了接近度的概念[49],获得相似的候选实体集。

2.2 扩展检索

借助于知识图谱网络,实体检索往往有着更大的自由度,其目标是从用户输入的实体对,甚至是自然语言推断用户的检索需求,而非之前的结构化语句和不包括关系的实体集合,这样扩展检索就能帮助不了解知识图谱以及不熟悉检索语言的用户为其提供检索服务。为此,扩展检索的目标已不再是满足用户基本的检索需求,关注用户多样化、多层次的检索需求成为研究的重点。

一般来说,实体扩展需要检索系统能够找到具有共同语义特征的实体,并给出候选实体集合。为此,计算实体之间的相似性,找到与用户查询相似的实体集合(或种子集合),是实现扩展检索的基本方法。如Y. Z. Sun等[50]在用户提出查询请求时,根据不同的侧重点,基于不同语义对实体之间的相似性进行计算,获得检索结果集。D. Mottin等[51]提出的EQ算法,基于子图同构和强模拟来构建等价关系,以返回相似的答案实体。N. Jayaram等[52]设计的GQBE也具有同样的功能,其方法是构建最大查询图和查询点阵,其效果要优于EQ。

对于用户采用自然语句作为检索式的问题,实现扩展检索主要采用基于规则和模板的方法。文献[53]采用将长检索式缩减为一个或多个子查询、赋予各实体权重、扩展查询的方法来完善信息、查询重构。文献[54]则采用基于模板的方式,将复杂的自然语言构成的问题生成结构化的SparQL语句,以解决基于句法无法适应复杂问题的情况。文献[55]针对问答系统,利用RDF三元组中的谓词约束其主体和客体的类型,以此生成查询图以涵盖不同类型的问题。文献[56]则从非结构化数据源中提取新的三元组,扩展知识图谱,并赋予新三元组与对应的原始三元组之间的可替代的权重,确保用户输入非专业术语时提供符合检索意图的信息。

识别检索用户的潜在需求,是扩展检索的另一个需要解决的问题。文献[57]基于路径的语义特征来描述各实体的共同方面,明确了用户想要返回的结果集的特征。推断用户的潜在检索意图是近年来的一种新的研究视角,通过分析用户输入实体的结构特征,综合各分支的相似实体,模拟用户的潜在兴趣[58],或判断用户输入的实体类型,计算信息增益,进而推断用户的检索意图[59]并给出检索结果。文献[60]则是利用实例驱动的方法来将语义相同但结构不同的子图进行等价匹配,实现检索扩展服务。在近年来的研究中,一

些研究者开始通过系统与用户之间进行交互来辅助用户完成扩展检索。文献[61]通过询问用户问题来消除歧义和模糊性,同时最小化交互成本,以保证用户的体验。文献[62]利用分面检索(facets search),允许用户继续点击检索实体在过滤器中的属性值,达到不断探索的目的。文献[63]设计的 Maverick(一种基于集束搜索框架开发的系统)可以发现实体中特殊实例的框架,有效地帮助用户探索独特的实体。此外,文献[64]提出的 JEDI(一种扩展了 SPARQL 引擎 Jena 的系统)对于用户了解特定实体同样有很大帮助,其可以提供知识图谱中两个节点如何实际连接的信息,包括路径和路径上的节点,而不仅仅显示两个节点是否连接。

随着检索逐步面向非专业用户,检索结果的多样性和多层次化的问题越来越受到重视。文献[65]设计的 X2Q(一种利用提供示例实现交互式的系统)基于用户输入的短语给用户提供查询建议,返回知识图谱中包含用户输入内容的正规实体名称以及实体的层次结构等。文献[66]利用最大边际相关性来避免返回结果同质。文献[67]则根据实体的上下文信息,包括实体在知识图谱中出现的频次以及实体类型的层次,来返回实体的合适类型。

2.3 推荐检索

实体推荐是指根据用户的检索请求提供相关对应实体的建议,以帮助用户更好地获得检索结果。与传统的推荐系统相比,实体推荐需要解决以下问题:①实体消歧,以满足用户的检索需求;②借助知识图谱,为用户提供跨领域的实体推荐;③为推荐的实体提供推荐理由;④使检索结果更直观、更具可读性。

实体推荐常用的方法是从知识图谱中分析具有直接联系的实体,利用实体间已知的关系,将存在直接关系的实体抽取出来作为相关实体集合,该方法被广泛应用[1]。对于实体间存在的间接关系,通过对实体间的相似度进行计算,判断两个实体的相关度[68]。一些研究通过知识图谱中实体所包含的属性以及实体描述文本间的相似性,来计算两个实体的相关程度。例如,有研究将实体对应的百科文章进行相似度计算,衡量两个实体的相关性[1]。也有采用潜在主题模型(latent Dirichlet Allocation,简称 LDA),通过对实体对应的描述文档进行建模并计算主题的相似程度,获得两个实体的相关度值[69]。

实体推荐需要给出一定的推荐理由,旨在解释被推荐实体与用户查询需求之间存在怎样的相关性,根据推荐理由,用户将判断是否进一步了解该实体。给出推荐理由的一般做法是利用知识图谱中已经存在的实体关系或利用预测得到的关系[70-71],作为实体间关系的注解,并推荐给用户。然而,这种方法存在信息量不足或注解缺乏语义性的问题。为了更好地解释两个实体间的给定关系,一些研究采用人工标注的方法,生成实体关系的解释句子[72]。该方法简单易行,但是需要为每一种实体关系人工标注一定数量的模板,因此不适合于大规模实体关系推荐任务。为了解决这个问题,文献[73]提出利用知识图谱获取特定的实体关系,自动生成解释句子模板。例如知识图谱中可能大量存在"人-作者-图书"的三元组关系,依据这种关系构建解释实体的关系图,然后根据这种关系自动学习出句子模板(如,"[图书]是[图书类型]的,作者是[人]。"),当给定新的三元组时,即可生成关系解释句子。然而,该方法受知识图谱中实体关系与实体属性覆盖率的限制,无法发现更多的实体间关系。

3 未来发展方向及面临的挑战

3.1 面向多元关系的实体关系推理

知识图谱是由实体和边构成的高度结构化的数据,这样的数据中蕴含了大量可以被机器所"理解"的语义信息。一元关系是指对应知识图谱中的实体,而二元关系则是指对应知识图谱中所有三元组的实体对。相对于一元和二元关系,多元关系的语义复杂程度更高,处理起来难度也更大。因此,实体关系的推理目前多集中在二元关系上,然而,将实体间的多元关系简化为二元关系,将带来大量的语义损失。一些研究已经发现,在 Freebase 中,有超过三分之一的实体存在多元关系[74]。

实现实体间多元关系的表达,意味着逻辑表达式的结构越复杂,灵活性越低,推理能力也会减弱。此外,为了实现多元关系的推理,还需要多元数据按照相对应的关系类型将对应的实体进行排序,降低了知识库的灵活性。因此,如何兼顾逻辑表达式与实体的推理能力是面向多元关系实体推理的研究方向,不仅需要在构建逻辑表达式时尽可能地减少实体间多元关系的信息结构,同时还需要确保推理过程的灵活。

3.2 检索结果可解释性

可解释性指系统是否能够对获得的检索结果给出语义解释。在实体扩展和实体推荐环节,可解释性显得尤为重要,因为系统给出的解释起到了"领域专家"与用户之间的桥梁作用,尤其是在一些需要专业知识才能对检索结果进行解读的领域。对检索结果进行解

释,是为了向用户说明系统提供答案的"依据",现有的基于知识图谱的检索技术中,大多基于实体的共同语义进行解释。然而,这种方法并没有考虑知识缺失的问题,也没有对间接语义进行计算,这就导致提供的共同语义存在局限性。

知识图谱存在知识的缺失,这使得对检索结果进行解释性面临着挑战。为解决这一问题,需要考虑:如何设定时间轴,解决实体间关系随着时间的变化而变化[72]以及如何对不存在直接关系的实体进行解释等问题。

3.3 与非结构化数据融合问题

实体搜索离不开丰富的数据源。相对于知识图谱的结构化特征,文本、视频、语音等非结构化数据在数量上远大于结构化数据集,将结构化的知识图谱与非结构数据相融合,可以从非结构化数据中将新出现的实体和新的关系补充到知识图谱中。

相较于结构化的数据,非结构化数据中,实体以及实体之间的关系更丰富。为此,基于非结构化数据的挖掘,可以将新出现的实体关系扩展到知识图谱中,提高查询的效果。一些研究已经开始对两者的融合进行探讨[75],但这方面研究仍处于起步阶段,缺乏通用的模型。

目前,神经网络算法的发展,使得在统一的空间中对不同类型(如文本、图像、语音等)数据进行表示成为可能,利用神经网络从非结构化数据中构建实体之间的关系,是值得探索的研究方向。

3.4 解决数据质量的问题

知识图谱的数据质量问题将影响检索的结果,数据质量问题主要是由知识缺失和知识错误造成的。

大多数知识图谱均存在知识缺失的问题,如 Wikipedia 无法覆盖所有实体。同时现有的知识图谱中也存在实体缺失三元组的问题[76],再如在 Freebase 中,93.8% 的人没有出生地信息,78.5% 的人没有国籍信息[77]。基于知识图谱的实体检索系统如未考虑知识缺失的问题,将影响最终结果的精度。解决知识缺失的问题可以通过以下两种途径:①知识补全,可通过基于规则的知识补全[78]或表示学习方法[79]利用知识库本身的知识对缺失部分进行补全,然而,该方法无法对知识图谱中没有的实体和关系进行补全;②知识推理,利用知识图谱中已有的知识去推理出新的事实(即隐含的知识),从而尽可能地对知识图谱进行补全[80]。

知识错误产生的原因比较多,可能是数据抽取过程中的错误,或实体之间存在语义的歧义,也可能是数据过时等。一些研究已经开始针对特定类型的错误进行监督,如文献[81]通过统计,可发现三元组中宾语是数值型的错误。然而,现有研究在错误知识发现问题上还缺乏统一的模型。使用人工监督移除不正确的事实元组是一种解决方法,但人工标注代价很高。也可通过知识推理方法,自动且高效地完成这一过程。

3.5 用户体验问题

现有的对于知识图谱检索结果的评价中广泛受到认可和使用的,仍是传统检索领域的指标,例如较为基础的精度和召回率[42,47]、检索效率(如查询时间[38])以及可扩展性[46,50]等。目前,一些研究已经开始关注用户体验的问题,如在用户使用知识图谱进行检索前后填写问卷,来收集用户反馈[24,30]。

结果多样性是提升用户体验的一个重要因素,因为其能避免结果冗余,所以文献[62]同时考虑相关性和多样性,以此来改良检索结果评价的指标。文献[82]则对 Google 和微软 Bing 的知识图谱进行了比较,主要对比了两者的实体类型覆盖范围、对列表检索以及自然语言检索的支持程度。然而,现有研究中,对于用户体验的关注多是分散和针对个案的,并没有形成公认通用的指标体系,这在未来是值得改进的地方。

3.6 跨领域知识图谱检索问题

尽管知识图谱本身的主题可能是属于一个或几个特定领域的,但是用户的检索需求可能并不局限于这些领域,甚至是知识图谱不包含的领域,这将会导致检索结果质量的下降。文献[83]设计的 RECAP,支持用户从多个以 RDF 进行编码的知识图谱中进行信息检索和知识发现,底层算法是较为通用的语法 SparQL。文献[30]则利用从多种来源文献中提取的科学元数据,构建知识图谱的 schema 层,从而实现跨学科、跨区域的科学资源检索。实体消歧和校正仍是不同知识图谱实体匹配的重要步骤,在这方面,文献[84]采用字嵌入的方式测量实体的相似性,然后进行实体匹配,大大降低了计算成本。

现有的跨领域知识图谱检索解决方案,大多是构建新的知识图谱,提供通用的数据特征,实现多领域知识图谱的相互链接。然而,这种方法受到知识图谱本身主题的限制,同时也需要大量成本和技术支持。

3.7 个性化推荐问题

检索对于用户来说是一项服务,为提升用户体验,个性化推荐是一项不错的功能。基于知识图谱的个性化推荐,主要体现在 2 个方面:①利用知识图谱向用户推荐感兴趣的内容,检索的过程就是推荐的过程,这方

面的个性化主要针对用户可能存在的需求,例如引文推荐[31],推特信息推荐[85]等;②在进行知识图谱检索时,对于不同的用户,在检索过程和返回结果上尽可能符合该用户的特征。

对于获取用户特征,可以仅根据检索式推断用户意图[33],这在初始缺少用户个人信息数据和检索行为数据时较多使用。而在有了一定的数据积累后,则可以利用历史数据,提升准确性,例如针对存在歧义的实体,可以根据用户过往的搜索主题进行消歧,或者对于检索式的扩展,可以给用户提供相似用户的检索式,以达到启发效果。

4 结语

知识图谱作为知识的一种结构化和语义化的表达方式,其强大的语义处理和互联组织能力,为智能化信息应用提供了基础。基于知识图谱的实体检索作为搜索引擎的主要方式,已经在智能搜索、智能问答、个性化推荐、内容分发等领域发挥重要的作用,也在教育科研、医疗、生物医疗以及需要进行大数据分析的一些行业展现出良好的应用前景。本文详细阐述了基于知识图谱的实体检索的定义、实现流程以及 3 种检索任务,并对实体检索应用场景与未来发展方向进行了总结。由于知识图谱的多学科结合的特征,需要知识库构建、自然语言理解、机器学习和数据挖掘等多方面研究的融合,共同推动该领域的研究和发展。

参考文献:

[1] HUANG J Z, DING S Q, WANG H F, et al. Learning to recommend related entities with serendipity for Web search users[J]. ACM transactions on Asian and low-resource language information processing, 2018, 17(3): 1-22.

[2] POUND J, MIKA P, ZARAGOZA H. Ad-Hoc object retrieval in the Web of data[C]// Proceedings of the 19th international conference on World Wide Web. New York: ACM Press, 2010: 771-780.

[3] SUCHANEK F M, KASNECI G, WEIKUM G. YAGO: a core of semantic knowledge[C]//Proceedings of the 16th international conference on World Wide Web. New York: ACM Press, 2007: 697-706.

[4] HOFFART J, SUCHANEK F M, BERBERICH K, et al. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia[J]. Artificial intelligence, 2013,194(1):28-61.

[5] MAHDISOLTANI F, BIEGA J, SUCHANEK F. YAGO3: a knowledge base from multilingual Wikipedias [EB/OL]. [2020-05-11]. https://hal-imt. archives-ouvertes. fr/hal-01699874/document.

[6] AUER S, BIZER C, KOBILAROV G, et al. DBpedia: a nucleus for a web of open data[C]//Proceedings of the 6th international semantic Web conference. Berlin: Springer-Verlag, 2007:722-735.

[7] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD international conference on management of data. New York: ACM Press, 2008: 1247-1250.

[8] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an architecture for never-ending language learning[C]//Proceedings of the 24th AAAI conference on artificial intelligence. Menlo Park: AAAI, 2010: 1306-1313.

[9] WU W, LI H, WANG H, et al. Probase: a probabilistic taxonomy for text understanding[C]//Proceedings of the 2012 ACM SIGMOD international conference on management of data. New York: ACM Press, 2012: 481-492.

[10] 徐增林,盛泳潘,贺丽荣,等. 知识图谱技术综述[J]. 电子科技大学学报,2016,45(4):589-606.

[11] 刘峤,李杨,段宏,等. 知识图谱构建技术综述[J]. 计算机研究与发展,2016,53(3):582-600.

[12] 林海伦,王元卓,贾岩涛,等. 面向网络大数据的知识融合方法综述[J]. 计算机学报,2017,40(1):1-27.

[13] QUILLIAN M R. Word concepts: a theory and simulation of some basic semantic capabilities[J]. Behavioral science, 1967,12(5): 410-430.

[14] SIMMONS R F. Natural language question answering systems: 1969 [J]. Communications of the ACM, 1970,13(1):15-30.

[15] BIZER C, HEATH T, BERNERS-LEE T. Linked data-the story so far[J]. International journal on semantic Web and information systems, 2009,5(3):1-12.

[16] DONG X L, GABRILOVICH E, HEITZ G, et al. Knowledge vault: a Web-scale approach to probabilistic knowledge fusion [C]//Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM Press, 2014: 601-610.

[17] LAMURIAS A, FERREIRA J D, CLARKE L A, et al. Generating a tolerogenic cell therapy knowledge graph from literature [J]. Frontiers in immunology, 2017,8(11):1-12.

[18] MUNGALL C J, MCMURRY J A, KÖHLER S, et al. The monarch initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species [J]. Nucleic acids research, 2017,45(1):712-722.

[19] YU T, LI J, YU Q, et al. Knowledge graph for TCM health preservation: design, construction, and applications[J]. Artificial intelligence in medicine, 2017,77(3):48-52.

[20] SHI L, LI S, YANG X, et al. Semantic health knowledge graph: semantic integration of heterogeneous medical knowledge and services [J]. BioMed research international, 2017, Article ID

2858423.
[21] FATHALLA S. Detecting human diseases relatedness[J]. International journal on semantic web and information systems, 2018,14(3):120－133.
[22] BAKAL G, TALARI P, KAKANI E V, et al. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations[J]. Journal of biomedical informatics, 2018,82(6):189－199.
[23] WANG X, WANG Y, GAO C, et al. Automatic diagnosis with efficient medical case searching based on evolving graphs[J]. IEEE access, 2018,6:53307－53318.
[24] CHI Y, YU C, QI X, et al. Knowledge management in healthcare sustainability: a smart healthy diet assistant in traditional Chinese medicine culture[J]. Sustainability, 2018,10(11):4197－4217.
[25] RUAN T, HUANG Y, LIU X, et al. QAnalysis: a question-answer driven analytic tool on knowledge graphs for leveraging electronic medical records for clinical research[J]. BMC medical informatics and decision making, 2019,19(1):82－94.
[26] ADAMS B, JANOWICZ K. Thematic signatures for cleansing and enriching place-related linked data[J]. International journal of geographical information science, 2015,29(4):556－579.
[27] FANG Y X, CHENG R, LI X D, et al. Effective community search over large spatial graphs[J]. Proceedings of the VLDB endowment, 2017,10(6):709－720.
[28] QIAO B, FANG K, CHEN Y, et al. Building thesaurus-based knowledge graph based on schema layer[J]. Cluster computing, 2017,20(1):81－91.
[29] GU L, XIA Y, YUAN X, et al. Research on the model for tobacco disease prevention and control based on case-based reasoning and knowledge graph[J]. Filomat, 2018,32(5):1947－1952.
[30] CHI Y, QIN Y, SONG R, et al. Knowledge graph in smart education: a case study of entrepreneurship scientific publication management[J]. Sustainability, 2018,10(4):995－1015.
[31] AYALA-GÓMEZ F, DARÓCZY B, BENCZÚR A, et al. Global citation recommendation using knowledge graphs[J]. Journal of intelligent & fuzzy systems, 2018,34(5):3089－3100.
[32] LIU D, LAI C. Mining group-based knowledge flows for sharing task knowledge[J]. Decision support systems, 2011,50(2):370－386.
[33] TOMEO P, FERNÁNDEZ-TOBÍAS I, CANTADOR I, et al. Addressing the cold start with positive-only feedback through semantic-based recommendations[J]. International journal of uncertainty, fuzziness and knowledge-based systems, 2017,25(S2):57－78.
[34] KARIM F, LYTRA I, MADER C, et al. DESERT: a continuous SPARQL query engine for on-demand query answering[J]. International journal of semantic computing, 2018,12(3):373－397.
[35] 张香玲，陈跃国，马登豪，等. 实体搜索综述[J]. 软件学报. 2017,28(6):1584－1605.
[36] NEUMAYER R, BALOG K, NORVAG K. On the modeling of entities for ad-hoc entity search in the Web of data[C]//Proceedings of the advances in information retrieval, 33rd European conference on IR research. Berlin: Springer-Verlag, 2012: 133－145.
[37] BALOG K, NØRVÅG K. On the use of semantic knowledge bases for temporally-aware entity retrieval[C]//Proceedings of the 5th workshop on exploiting semantic annotations in information retrieval. New York: ACM Press, 2012: 1－2.
[38] LASHKARI F, ENSAN F, BAGHERI E, et al. Efficient indexing for semantic search[J]. Expert systems with applications, 2017,73(5):92－114.
[39] KATIB A, SLAVOV V, RAO P. RIQ: fast processing of SPARQL queries on RDF quadruples[J]. Journal of Web semantics, 2016,37－38(S):90－111.
[40] HAO Y, LI G, YUAN P, et al. An association-oriented partitioning approach for streaming graph query[EB/OL]. [2020－05－11]. http://downloads.hindawi.com/journals/sp/2017/2573592.pdf.
[41] ZHENG W, LIAN X, ZOU L, et al. Online subgraph skyline analysis over knowledge graphs[J]. IEEE transactions on knowledge and data engineering, 2016,28(7):1805－1819.
[42] SONG Q, WU Y, LIN P, et al. Mining summaries for knowledge graph search[J]. IEEE transactions on knowledge and data engineering, 2018,30(10):1887－1900.
[43] LI J, WU F D, YANG X, et al. APPQKW: an efficient algorithm for processing path queries on the knowledge graph under ecosystem[J]. Ekoloji, 2019,28(107):2031－2040.
[44] YUAN Y, WANG G, CHEN L, et al. Efficient pattern matching on big uncertain graphs[J]. Information sciences, 2016,339(4):369－394.
[45] LIN X, PENG Y, CHOI B, et al. Human-powered data cleaning for probabilistic reachability queries on uncertain graphs[J]. IEEE transactions on knowledge and data engineering, 2017,29(7):1452－1465.
[46] HE L, LIU B, LI G, et al. Knowledge base completion by variational bayesian neural tensor decomposition[J]. Cognitive computation, 2018,10(6):1075－1084.
[47] KLIEGR T, ZAMAZAL O. LHD 2.0: a text mining approach to typing entities in knowledge graphs[J]. Journal of Web semantics, 2016,39(8):47－61.
[48] CAO T H, HUYNH D T. Subsumption degrees between entity types and names for approximate knowledge retrieval[J]. International journal of uncertainty fuzziness and knowledge-based systems, 2007,15(1):21－42.
[49] JIN J, LUO J, KHEMMARAT S, et al. Querying Web-scale knowledge graphs through effective pruning of search space[J]. IEEE transactions on parallel and distributed systems, 2017,28(8):2342－2356.
[50] SUN Y Z, HAN J W, YAN X F, et al. PathSim: meta path-based top-K similarity search in heterogeneous information networks[J].

PVLDB, 2011,4(11):992 – 1003.
[51] MOTTIN D, LISSANDRINI M, VELEGRAKIS Y, et al. Exemplar queries: a new way of searching[J]. The VLDB journal, 2016,25(6):741 – 765.
[52] JAYARAM N, KHAN A, LI C, et al. Querying knowledge graphs by example entity tuples[J]. IEEE transactions on knowledge and data engineering, 2015,27(10):2797 – 2811.
[53] GUPTA M, BENDERSKY M. Information retrieval with verbose queries[J]. Foundations and trends in information retrieval, 2015, 9(3/4):209 – 354.
[54] ZHENG W, YU J X, ZOU L, et al. Question answering over knowledge graphs [J]. Proceedings of the VLDB endowment, 2018,11(11):1373 – 1386.
[55] SHIN S, JIN X, JUNG J, et al. Predicate constraints based question answering over knowledge graph[J]. Information processing & management, 2019,56(3):445 – 462.
[56] YAHYA M, BERBERICH K, RAMANATH M, et al. Exploratory querying of extended knowledge graphs[J]. Proceedings of the VLDB endowment, 2016,9(13):1521 – 1524.
[57] CHEN J, CHEN Y, ZHANG X, et al. Entity set expansion with semantic features of knowledge graphs[J]. Journal of Web semantics, 2018,52/53(10):33 – 44.
[58] METZGER S, SCHENKEL R, SYDOW M. QBEES: query-by-example entity search in semantic knowledge graphs based on maximal aspects, diversity-awareness and relaxation[J]. Journal of intelligent information systems, 2017,49(3):333 – 366.
[59] SHAN Y, LI M, CHEN Y. Constructing target-aware results for keyword search on knowledge graphs[J]. Data & knowledge engineering, 2017,110(7):1 – 23.
[60] ZHENG W G, ZOU L, et al. Semantic SPARQL similarity search over RDF knowledge graphs[J]. Proceedings of the VLDB endowment, 2016,9(11):840 – 851.
[61] ZHENG W, CHENG H, YU J X, et al. Interactive natural language question answering over knowledge graphs[J]. Information sciences, 2019,481(5):141 – 159.
[62] ARENAS M, CUENCA GRAU B, KHARLAMOV E, et al. Faceted search over RDF-based knowledge graphs[J]. Journal of Web semantics, 2016,37 – 38(3):55 – 74.
[63] ZHANG G, LI C. Maverick[J]. Proceedings of the VLDB endowment, 2018,11(12):1934 – 1937.
[64] AEBELOE C, MONTOYA G, SETTY V, et al. Discovering diversified paths in knowledge bases[J]. Proceedings of the VLDB endowment, 2018,11(12):2002 – 2005.
[65] LISSANDRINI M, MOTTIN D, VELEGRAKIS Y, et al. X(2)Q: Your personal example-based graph explorer[J]. Proceedings of the VLDB endowment, 2018,11(12):2026 – 2029.
[66] ARNAOUT H, ELBASSUONI S. Effective searching of RDF knowledge graphs[J]. Journal of Web semantics, 2018,48(1):66 – 84.
[67] TONON A, CATASTA M, PROKOFYEV R, et al. Contextualized ranking of entity types based on knowledge graphs[J]. Journal of Web semantics, 2016,37/38(3):170 – 183.
[68] BIN B I, HAO M A, HSU BO-JUNE (PAUL), et al. Learning to recommend related entities to search users[C]//Proceedings of the 8th ACM international conference on Web search and data mining. New York: Association for Computing Machinery, 2015: 139 – 148.
[69] BLEI M D, NG Y A, JORDAN I M. Latent dirichlet allocation [J]. Journal of machine learning research, 2003, 3(4/5):993 – 1022.
[70] LIN Y K, LIU Z Y, SUN M S, et al. Learning entity and relation embeddings for knowledge graph completion[EB/OL]. [2020 – 05 – 11]. https://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9571/9523.
[71] JI G L, LIU K, HE S Z, et al. Knowledge graph completion with adaptive sparse transfer matrix[EB/OL]. [2020 – 05 – 11]. https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11982/11693.
[72] ALTHOFF T, DONG X L, MURPHY K, et al. TimeMachine: timeline generation for knowledge-base entities[C]//Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining. New York: Association for Computing Machinery, 2015: 19 – 28.
[73] VOSKARIDES N, MEIJ E, DE RIJKE M. Generating descriptions of entity relationships[C]// Proceedings of 39th European Conference on IR Research. Switzerland: Springer Nature, 2017: 317 – 330.
[74] WEN J, LI J, MAO Y, et al. On the representation and embedding of knowledge bases beyond binary relations[EB/OL]. [2020 – 05 – 11]. https://arxiv.org/pdf/1604.08642.pdf.
[75] LEE J Y, MIN J K, OH A, et al. Effective ranking and search techniques for Web resources considering semantic relationships [J]. Information processing and management, 2014,50(1):132 – 155.
[76] SUCHANEK F M, GROSS-AMBLARD D, ABITEBOUL S. Watermarking for ontologies[C]// The Semantic Web - ISWC 2011. Berlin: Springer-Verlag, 2011: 697 – 713.
[77] MIN B, GRISHMAN R, WAN L, et al. Distant supervision for relation extraction with an incomplete knowledge base[C]//Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics: human language technologies. Stroudsburg: ACL, 2013: 777 – 782.
[78] CHEN Y, GOLDBERG S L, WANG D Z, et al. Ontological pathfinding[C]//Proceedings of the 2016 ACM SIGMOD international conference on management of data. New York: ACM Press, 2016: 835 – 846.
[79] LIN Y K, LIU Z Y, SUN M S, et al. Learning entity and relation embeddings for knowledge graph completion[C]//Proceedings of

the 29th conference on artificial intelligence. Palo Alto: AAAI, 2015: 2181－2187.
[80] GARDNER M, MITCHELL T M. Efficient and expressive knowledge base completion using subgraph feature extraction[C]//Proceedings of the conference on empirical methods in natural language processing. Stroudsburg: Association for Computational Linguistics, 2015: 1488－1498.
[81] WIENAND D, PAULHEIM H. Detecting incorrect numerical data in DBpedia[C]//Proceedings of the semantic Web: trends and challenges, 11th internationl conference. Berlin: Springer-Verlag, 2014: 504－518.
[82] UYAR A, ALIYU F M. Evaluating search features of Google knowledge graph and Bing satori[J]. Online information review, 2015,39(2):197－213.
[83] FIONDA V, PIRRÒ G. Explaining and querying knowledge graphs by relatedness[J]. Proceedings of the VLDB endowment, 2017,10(12):1913－1916.
[84] YAN J, XU C, LI N, et al. Optimizing model parameter for entity summarization across knowledge graphs[J]. Journal of combinatorial optimization, 2019,37(1):293－318.
[85] PLA KARIDI D, STAVRAKAS Y, VASSILIOU Y. Tweet and followee personalized recommendations based on knowledge graphs[J]. Journal of ambient intelligence and humanized computing, 2018,9(6):2035－2049.

作者贡献说明：

阮光册:论文总体框架设计,论文修改;
樊宇航:资料收集,论文初稿撰写;
夏磊:论文修改。

A Review of the Application of Knowledge Graph in Entity Retrieval

Ruan Guangce[1] Fan Yuhang[1] Xia Lei[2]
[1] Department of Information Management, Faculty of Economics and Management, East China Normal University, Shanghai 200241
[2] Shanghai Library, Shanghai 200031

Abstract: [Purpose/significance] To sort out the research context and key points of entity retrieval based on knowledge graph, and explore the future development direction of this field. [Result/conclusion] This paper firstly gave the formal definition, the implementation path and main data sources of entity retrieval on knowledge graph. Then, according to the retrieval task, the application of entity retrieval was divided into match retrieval, extended retrieval and recommended recommendation, and the implementation methods were summarized. [Result/conclusion] With the development of the application, the research of entity retrieval based on knowledge graph began to focus on how to improve the user's retrieval experience and provide a variety of retrieval results. The future research will be carried out on the interpretability of retrieval results, cross domain knowledge graph retrieval and so on.

Keywords: entity retrieval knowledge graph information retrieval